Assessing item fairness in students’ evaluation of teaching based on students’ academic college using measurement invariance analysis

Mahmoud AlQuraan
Department of Counseling and Educational Psychology, Yarmouk University, Irbid, Jordan and
Al Ain University, Al Ain, United Arab Emirates

Abstract
Purpose – This study aims at assessing item fairness in students’ evaluation of teaching based on students’ academic college using measurement invariance analysis (MI).
Design/methodology/approach – The sample of this study consists of 17,270 undergraduate students from 12 different academic colleges. SET survey consists of 20 Likert-type items distributed to four factors: planning, instruction, management and assessment was used to collect the data. The Lavaan R package with confirmatory factor analysis (CFA) was used to evaluate measurement invariance (MI). Four models of CFA were investigated and assessed: the configural model, the metric model, the scalar model and the residual invariance model. ANOVA was used to test the differences in SET according to academic colleges.
Findings – MI analysis showed that the four levels of MI models are supported. ANOVA test showed that means of SET total scores are statistically different according to students’ academic colleges. College of “Education” has the highest SET mean (88.64 out of 100), and all the differences between the College of Education’s SET mean and other colleges’ SET means are statistically significant.
Practical implications – The study recommends that higher education institutions test the MI of SET according to academic colleges and then use colleges with the highest SET at the university level as internal benchmarking to develop and enhance their teaching practices.
Originality/value – This study is probably the only study that tested MI according to students’ colleges before testing the differences between colleges in SET. If MI is not supported, then the comparisons between academic colleges are not applicable.
Keywords Academic colleges, Measurement invariance (MI), Students evaluation of teaching (SET)
Paper type Research paper

Introduction
Student evaluation of teaching (SET) is a common practice in higher education as a measure of institutional effectiveness (Cunningham et al., 2023; Ching, 2018). It is used for critical decisions related to faculty members, such as retention, tenure and promotion (Uttl, 2021). Accreditation agencies use SET as key performance indicators at the university and program levels. Most universities use a unified survey to collect SET data (Alquraan, 2019). This survey is distributed at the end of the semester despite students’ academic college. Marsh (2007) stated that SET is multidimensional, reliable and valid against a variety of indicators of effective teaching, unaffected by a variety of potential biases, and useful for faculty, students and administrators. So greater attention should be given to SET surveys’ consequential validity to achieve the purposes of SET as no agreed-upon definition of “effective teaching” or any single, all-embracing criterion (Cashin, 2003).

This study was conducted during the sabbatical leave from Yarmouk University, Jordan.
Students evaluation of teaching

Tools have been developed to measure teaching, courses and programs' effectiveness, and they have been used in almost every university. Based on SET data, reports are generated across instructors, departments and colleges and utilized as evidence of teaching effectiveness and professional development decisions (Sproule, 2000).

There is evidence that SETs are valid sources of information about learning environment productivity, usefulness or worthiness (Cook et al., 2022), and they are valuable feedback that teachers can use to develop their teaching practices (Marsh et al., 1992). Ramsden’s (1991) study showed that students are reliable, verifiable and useful sources of determining the teaching quality of academic units in systems of higher education that are based on British models. In a meta-analysis study conducted by Cohen (1981), a strong support for the validity of student ratings as measures of teaching effectiveness was found.

Many factors (independent variable) influence SET (dependent variable): Grades or expected grades (Griffin et al., 2014), gender (Sigurdardottir et al., 2023), teachers’ characteristics (Tran and Do, 2022), classroom size and response rate (Badri et al., 2006), course difficulty (Addison et al., 2006), course level (Santhanam and Hicks, 2002), course type (Beran and Violato, 2005), general versus specific education (Ting, 2000) and course syllabus tone (Harnish and Bridges, 2011).

Using data for 343 courses from 11,203 surveys in a Korean four-year private university, Choi and Kim (2014) found that 96% of the variance of monotonic response patterns could be explained by student characteristics, such as gender, academic year, major, grade point average and perceptions about course difficulty. The findings of this study suggest that Korean higher education institutions need to have proper surveys to encourage students to participate actively in SET for the best and proper use of reports generated based on SET results.

In a meta-analysis study, Uttl et al. (2017) analyzed published meta-analyses of the multisection studies about SET and found that there is an artifact of small sample-sized studies and publication bias. Whereas the small sample-sized studies showed a large and moderate correlation, the large sample-sized studies showed no or only minimal correlation between SET ratings and learning. Their meta-analysis of all multisection studies revealed no significant correlations between the SET and learning. These results yield the authors to suggest that institutions focused on student learning and career success may want to ignore SET ratings as a measure of faculty’s teaching effectiveness.

Also, students’ academic discipline is one of the factors that have a significant effect on students’ rating of teaching (Neumann, 2001; Chen and Watkins, 2010; Basow and Montgomery, 2005), and the wording of SET’s survey items could be one of the reasons behind the effect of students’ discipline on SETs as shown by Anders et al. (2016). This implies that some items might be worded to be understood in a different way according to students’ academic college. In other words, some items’ content might give advantages or disadvantages to a particular field of study over another. In psychometric terms, students’ endorsement of an option (response) on a Likert scale item could be influenced by students’ academic college rather than what this survey measures.

Jones et al. (2022) found that students rated easier courses lower when controlling for motivational climate and the demographical composition of the class. Constantinou and Wijnen-Meijer (2022) provided insight that when utilizing SET, it is important to be aware of important factors that may influence SET data, such as low attention and lack of time from students, biases due to course difficulty and grades, students’ motivation as well as gender biases.

Sullivan et al’s (2023) review identified a variety of research related to students’ motivations, perceptions and opinions about SET. This affords direction and suggestions for educational administrations to refine processes and structures for collecting SET data that
will be meaningful. They found that perceptions, knowledge and attitudes about SET were identified as important factors in motivating students to engage and complete SETs, mainly if surveys were easy to interpret, their time for completing them was short and they were convinced that their responses would be valued.

The previously mentioned studies show that SET might be affected by variables related to students or the courses. This could be threatening of the construct validity of SET surveys and might encourage to ignore using SET (Uttl et al., 2017). The current study investigates item fairness in SET surveys utilizing data collected from a major university in the country of Jordan. It links students’ academic college and SET item fairness, which threatens the survey’s validity and reliability. The probability of endorsing an option or point in a rating scale item should be determined by the latent trait (e.g. teaching and learning effectiveness) measured by the survey that said item comes from.

SET survey items – teaching practices – should be perceived and understood in the same way despite students’ academic college (Schoot et al., 2012). If students from different colleges understand SET survey items differently, then the interpretations and uses of SET results may not be valid. Psychometrically, examining the validity of the SET theoretical construct is called testing measurement invariance of the SET survey (Dimitrov, 2010). This is the main objective of the current study.

Measurement invariance
Measurement invariance (MI) of SET survey items indicates how these items (teaching practices) are related to students’ college. In the context of the current study, MI is established if the quantitative relationships of teaching practices (survey items) to the SET theoretical construct measured by these items are identical across students’ colleges. If MI is established, then the students who filled the self-reported surveys across all colleges interpret survey items and the measured theoretical construct in the same way, and the comparisons between colleges can be made (Krammer et al., 2021; Schoot et al., 2012).

MI is assumed when measurement models are used to know the differences in a measured construct between groups (Clark and Donnellan, 2021). Psychometric scholars use quantitative methods to evaluate MI. Asparouhov and Muthén (2014) recommended multi-group confirmatory factor analysis (MGCFA) to evaluate MI. MGCFA tests the hypothesis that “a given theoretical model fits well to the data across all groups under investigation.”

To use MGCFA to test MI, several steps were presented by Schoot et al. (2012). These steps start with a confirmatory factor analysis (CFA) model for each group, so the construct validity of the latent variable (measured attribute) can be evaluated using a number of model fit indices. This stage tests and measures configural invariance (structural equivalence). This stage tests the item loadings across groups (colleges). The model in this stage has no constraint across groups other than the loadings between survey items and the latent variable. The stage helps researchers to test whether the basic structure is supported or whether the pattern of loadings of survey items on the measured theoretical construct differs among the groups. Violating configural invariance indicates that students from different colleges perceive the SET items (teaching practice) differently. In the second stage, only the factor loadings are set to be equal across groups, while the intercepts are allowed to be free across groups. This stage tests the metric invariance (factor loadings (slopes) are the same across all groups), which provides researchers with information on whether responders from different groups have the same meaning of the latent variable under investigation. Also, the model needs to run where the intercepts are set to be equal across groups while the factor loadings of the survey items are allowed to be free among groups. This stage tests whether the items’ intercepts are equal across groups or not. In the third stage, a model where the
loadings and intercepts are constrained to be equal. This stage tests scalar invariance to know whether the latent variable’s meaning and the underlying items (intercepts) levels are equal across groups.

Model fit is evaluated using a number of global fit indices. Three model fit indices were recommended by Rutkowski and Svetina (2014) and used in this study. Tucker–Lewis Index (TLI), Comparative Fit Index (CFI) and standardized root mean squared residual (SRMR). CFI and TLI compare the fit of targeted models, and they should be > 0.90. SRMR is the square root of the difference between the residuals of the sample covariance matrix and the hypothesized model, and it should be <0.08. For model comparisons, differences of the fit indices are insignificant if the CFI and TLI decreased less than 0.010, and the SRMR decreased less than 0.03 (Chen, 2007).

The current study is trying to answer the following main research question, “What are the sources of variation in SET survey construct validity according to students’ academic college?” using measurement invariance analysis as it one of the best statistical and psychometric analyses that could be used to achieve this goal. MI is required whenever SET scores are compared. It has an important role in the context of SET survey construction and validation. Thus MI should be established to ensure that the SET are comparable across subgroups (Colleges in the context of this study).

The rationale for the current study
Reviewing the previous literature shows that research studies focus on the factors influencing SET. These research efforts have analyzed SET as a dependent variable and the other factors were analyzed as independent variables. There is a disagreement among the previous studies about the relationships between SET total score and other independent factors. This disagreement could make educators and researchers question the measurement equivalency of SET according to the groups forming the independent variables. This study is trying to contribute to this issue by investigating item fairness in SET surveys. Investigation of item unfairness means discovering whether the survey items have the same meaning for responders of different subgroups of the population. Item unfairness might be the reason behind this disagreement about the relationship between students’ rating of teaching and the studied factors as violating measurement invariance of SET survey based on some factors means studying the relationships between SET and these factors are inappropriate and unfair.

Methodology
The instrument
This study used the data collected by Yarmouk University. Yarmouk University uses a SET survey that is approved by the authorized councils at that university to monitor teaching effectiveness and quality. The university uses a self-reported SET survey to evaluate teaching effectiveness which is distributed at the end of each semester. This survey consists of 20 Likert-type items distributed to four factors: planning, instruction, management and assessment teaching practices; each factor is measured by five items. The content validity of the SET survey was evaluated by panel experts in the related fields. The reliability and validity of the SET survey were evaluated using Cronbach’s alpha, CFA, and exploratory factor analysis (EFA). The Cronbach’s alpha as a reliability index was found to be 0.94. The four-factor CFA model analysis presented in detail in the results of this study provided evidence of the SET survey construct validity. The EFA showed that a common factor explains 85% of the total variance, which is considered evidence of the SET survey validity.
Sample
The sample of this study consists of 17,270 students distributed to 12 academic colleges, as shown in Table 1. This sample is part of a large dataset consisting of 166,938 responses. The data was collected by the university quality assurance center for monitoring and assessment purposes and given to the author to conduct the study. The number of students in this university is close to 43,000 students, and every student evaluates all the teachers who teach this student that semester. So every student filled out four surveys on average. Therefore, the researcher selected the SET data for the courses offered simultaneously to ensure that the sample was an independent sample and that every student completed only one survey.

Data analysis
To evaluate MI, the Lavaan R package with confirmatory factor analysis (CFA) was used (Rosseel, 2012). Four models of CFA were investigated and assessed. The first model was the configural model or invariance of the model form. MGCFA was run separately for each academic college. This model has no constraint across colleges other than the loadings between survey items and the latent variable. The second model was the metric model or equivalence of the item loadings on the factors. In this model, the equivalence of factor loadings (slopes) is set to be the same across all colleges to test whether students from different colleges attribute the same meaning to students’ evaluation of teaching practices. The third model is the scalar model or equivalence of item intercepts. Scalar invariance is tested by constraining the item intercepts to be equivalent across all academic colleges retaining the constraints applied in the metric invariance model. The fourth model is the residual invariance model, which constrains the residual variances to be equal across colleges. Residual invariance indicates that the variance of the item that is not shared with the factor and the measurement error is similar across colleges. Residual invariance is an assumption that is difficult to achieve in practice because it assumes that any unexplained variance not due to the factor is also equivalent across colleges (see Table 2).

Results
First, CFA was conducted for all colleges (Model_ 0 all). Then, the four MI models were assessed using MGCFA. The fit indices of these models are presented in Table 3. Overall and based on Chen’s (2007) thresholds values, the content of Table 3 shows that Model 0 (All

<table>
<thead>
<tr>
<th>College</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>1397</td>
<td>8.1</td>
</tr>
<tr>
<td>Archaeology and anthropology</td>
<td>845</td>
<td>4.9</td>
</tr>
<tr>
<td>Al-Sharee’a and Islamic studies</td>
<td>2,255</td>
<td>13.1</td>
</tr>
<tr>
<td>Economics and administration science</td>
<td>1,169</td>
<td>6.8</td>
</tr>
<tr>
<td>Engineering</td>
<td>298</td>
<td>1.7</td>
</tr>
<tr>
<td>Arts</td>
<td>2,953</td>
<td>17.1</td>
</tr>
<tr>
<td>Mass communication</td>
<td>168</td>
<td>1.0</td>
</tr>
<tr>
<td>Education</td>
<td>2,257</td>
<td>13.1</td>
</tr>
<tr>
<td>Fine arts</td>
<td>1,491</td>
<td>8.6</td>
</tr>
<tr>
<td>Physical education</td>
<td>1,402</td>
<td>8.1</td>
</tr>
<tr>
<td>Law</td>
<td>1,890</td>
<td>10.9</td>
</tr>
<tr>
<td>Tourism and hotel management</td>
<td>1,145</td>
<td>6.6</td>
</tr>
<tr>
<td>Total</td>
<td>17,270</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Table 1. Sample distribution according to students’ academic college

Source(s): Authors’ work
Colleges) has a good (CFIAll = .983; TLIAll = .981; SRMRAll = .007; RMSEAAll = .063). These results again provide evidence of SET unidimensionality, indicating that students’ evaluation of teaching practice reflects a unitary construct.

Since the fit indices of the configural model, which is presented in Table 3, reached the accepted level, the same factorial structure holds across all colleges (CFIConfigural = .973; TLIConfigural = .968; SRMRConfigural = .010; RMSEAConfigural = .081). The configural invariance model is used as a reference model to compare the fit of metric invariance.

The metric invariance model is a constrained version of the configural model where the factor loadings are assumed to be equal across students’ academic colleges, but the intercepts are allowed to be free. This stage tested whether the meaning of the teaching practices (intercepts) levels are equal across different academic colleges. The results of testing the metric model are shown in Table 3. Testing metric invariance involves comparing the configural model against the metric model using CFI and SRMR change (Δ). For model comparisons, fit indices change is insignificant if the CFI changes to less than 0.010 and the SRMR changes to more than 0.03 (Chen, 2007). The results presented in Table 3 show that the fit indices are not significantly worse, which suggests that constraining the loadings of the items across students’ academic colleges does not statistically affect the model fit, and metric invariance is supported (see Figures 1 and 2).

Since testing the metric model showed that the meaning of the levels of the teaching practices included in the SET survey (intercepts) are the same students’ academic colleges, the scalar model was run. The global fit indices of scalar invariance are also shown in Table 3. Δ CFI and Δ SRMR do not reach the cut-off values suggested by Chen (2007) (0.010 for CFI and 0.03 for SRMR). Testing the scalar invariance indicates that mean differences in SET capture all mean differences in the shared variance of the teaching practices. It means that constraining the item intercepts across academic colleges does not statistically affect the model fit indices, and scalar invariance is supported.

The results presented in Table 3 showed a non-significant change in CFI and SRMR between the metric and scalar models, which means that the residuals invariance assumption

<table>
<thead>
<tr>
<th>MI levels</th>
<th>Item loadings</th>
<th>Item thresholds</th>
<th>Item residuals</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural model</td>
<td>Free</td>
<td>Free</td>
<td>Free</td>
<td>No group comparisons</td>
</tr>
<tr>
<td>Metric model</td>
<td>Equal</td>
<td>Free</td>
<td>Free</td>
<td>Variance and covariance can be compared</td>
</tr>
<tr>
<td>Scalar model</td>
<td>Equal</td>
<td>Equal</td>
<td>Free</td>
<td>Means, variances and covariance can be compared</td>
</tr>
<tr>
<td>Residual model</td>
<td>Equal</td>
<td>Equal</td>
<td>Equal</td>
<td>True group differences are the only source of means differences</td>
</tr>
</tbody>
</table>

**Table 2. Summary of testing MI levels and implications**

Source(s): Authors’ work

<table>
<thead>
<tr>
<th>Model</th>
<th>χ²</th>
<th>df</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>SRMR</th>
<th>ΔCFI</th>
<th>ΔSRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 0 – CFA all</td>
<td>11254.091</td>
<td>164</td>
<td>0.983</td>
<td>0.981</td>
<td>0.063</td>
<td>0.007</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model 1 – configural</td>
<td>20421.638</td>
<td>1968</td>
<td>0.973</td>
<td>0.968</td>
<td>0.081</td>
<td>0.010</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Model 2 – metric</td>
<td>20649.342</td>
<td>2144</td>
<td>0.973</td>
<td>0.971</td>
<td>0.077</td>
<td>0.013</td>
<td>0.000</td>
<td>–0.003</td>
</tr>
<tr>
<td>Model 3 – scalar</td>
<td>20983.413</td>
<td>2320</td>
<td>0.972</td>
<td>0.973</td>
<td>0.075</td>
<td>0.013</td>
<td>-0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>Model 4 – residual</td>
<td>37843.197</td>
<td>2540</td>
<td>0.948</td>
<td>0.953</td>
<td>0.098</td>
<td>0.015</td>
<td>-0.006</td>
<td>0.002</td>
</tr>
</tbody>
</table>

**Table 3. Fit indices of the MGCFA for evaluating MI models**

Source(s): Authors’ work
can be tested and evaluated. The global fit indices of residual invariance are also shown in Table 3. Δ CFI and Δ SRMR do not reach the cut-off values suggested by Chen (2007) (0.010 for CFI and 0.03 for SRMR). Testing for residual invariance or equivalence of item residuals of metric and scalar invariant items indicates that the sum of specific variance and error variance is similar across academic colleges. Testing residual invariance is not a prerequisite for testing mean differences of SET scores between colleges because the residuals are not part of SET total scores, so invariance of the item residuals is inconsequential to the interpretation of SET mean differences (Vandenberg and Lance, 2000). Therefore, one way analysis of variance (ANOVA) was used to investigate the means differences in SET according to students’ academic college. The results are presented in Tables 4–6 and Figure 3. Table 4 shows that there are statistical differences among the means SET total scores according to

**Figure 1.** CFI and TLI for the four MI models: Higher the better

**Figure 2.** RMSEA and SRMR for the four MI models: Lower the better

Source(s): Authors’ work
students’ academic college. Table 5 shows that students of the College of “Physical Education” has the lowest SET mean (74.27), while students of the Colleges of “Mass Communication” and “Education” have the highest SET mean.

Table 6 shows the results of the Scheffe post hoc test of pairwise comparisons. The table shows that the differences are mainly between the three groups of colleges. The first group, low SET level, is the Colleges of “Science,” “Engineering” and “Physical Education.” The means of the SET total score for these colleges are very close to 75 out of 100. The second group of colleges, medium SET level, is colleges of “Archaeology and Anthropology,” “Al-Sharee’a and Islamic Studies,” “Economics and Administration Science,” “Arts,” “Fine Arts,” “Law” and “Tourism and Hotel Management.” The means of the SET total score for these colleges range from 80 to 85 out of 100. The third group, high SET level, is colleges of “Mass Communication” and “Education.” These two colleges have means of the SET total score very close to 89 out of 100.

Summary and implications
The results of using measurement invariance analyses in this study showed that the four-level measurement equivalency is supported. Invariance at the configural level in the SET survey, which is supported in this study, indicates that the basic organization of the SET survey (four-factor Model with five loadings of the teaching practices) is supported across students’ academic colleges. This means that the pattern of loadings of the teaching practices measured by the four-factor SET (planning, instruction, management and assessment) is the same across colleges and suggests that the overall factor structure holds up similarly for all academic colleges. This result means that there is no need to refine the SET survey or omit any item or teaching practices.
Metric invariance, which is supported in this study, means that each teaching practice in the SET survey contributes to the measurement of teaching effectiveness to a similar level across all students’ academic colleges. Scalar invariance, which is supported in this study, suggests that mean differences in the SET total score capture mean differences in the common variance of the teaching practices. Residual invariance, which is supported in this study, indicates that the variance of the teaching practices that are not common with the SET factor and error variance is the same across students’ academic colleges. Since the four measurement

<table>
<thead>
<tr>
<th>Mean</th>
<th>Academic college</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.22</td>
<td>1. Science</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84.22</td>
<td>2. Archaeology and Anthropology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81.73</td>
<td>3. Al-Shar’ee’a and Islamic Studies</td>
<td></td>
<td></td>
<td>N.S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>84.24</td>
<td>4. Economics and Administration Science</td>
<td></td>
<td></td>
<td>N.S</td>
<td>N.S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>75.30</td>
<td>5. Engineering</td>
<td>N.S</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>80.35</td>
<td>6. Arts</td>
<td>*</td>
<td>*</td>
<td>N.S</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88.64</td>
<td>7. Mass Communication</td>
<td>*</td>
<td>N.S</td>
<td>*</td>
<td>N.S</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88.51</td>
<td>8. Education</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82.68</td>
<td>9. Fine Arts</td>
<td>*</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>74.27</td>
<td>10. Physical Education</td>
<td>N.S</td>
<td>*</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81.51</td>
<td>11. Law</td>
<td>*</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>*</td>
<td>*</td>
<td>N.S</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>83.74</td>
<td>12. Tourism and Hotel Management</td>
<td>*</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>*</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td>N.S</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note(s):** *The difference between the two means is statistically significant N.S. The difference between the two means is not statistically significant

**Source(s):** Authors’ work

Table 6. Scheffe post hoc test results

**Figure 3.**
SET means across academic colleges

Source(s): Authors’ work
invariance models are supported, there is evidence that the SET survey does not have unfair items, and SET can be compared according to students’ academic colleges (DeShon, 2004). When residual invariance is supported as in this study, true group differences are the only source of means differences (Geiser et al., 2014). This means that the existence of mean differences among colleges in SET are true differences.

Since the measurement invariance of the SET survey according to students’ academic colleges is supported, ANOVA was used to test means differences in SET total scores among colleges. The results showed statistical differences, and the means of SET for the 12 colleges ranged from 74.27 to 88.64 out of 100. College of “Education” has the highest SET mean (88.64), and all the differences between the College of Education’s SET mean, and other colleges’ SET means are statistically significant. Marsh (2007) stated that most SET surveys developed around the concept of effective teaching. These teaching practices included in SET survey should be well known for the faculty member at college of education as a part of their qualifications. This might be the reason behind getting the highest SET mean.

Best teaching practices are the target of higher education institutions to enhance students’ learning (Zimmerman et al., 2020). The current study’s results can help establish measures of quality tools and processes used by institutions towards continuous improvement and benchmarking of the quality of teaching practices. College of Education, in the context of this study, can be used as a successful internal benchmark, doing an internal review of the teaching practices used by the College of Education faculty members could be used as a model for the faculty members in other colleges to develop their teaching practices.

It should be taken into account that the results of this study are based on 20 items SET survey, students evaluation of teaching survey, and one university experience. Most universities use more than one survey to measure students point of views such as course evaluation survey.

In general, this study suggests the following to benefit from SET:

1. Using measurement invariance analysis to assure SET construct validity. This might require deleting some items from the survey to make sure that true college differences are the only source of means differences in the dataset.

2. Doing more research to investigate the measurement invariance of SET survey based on other factors that previous research showed that SET is affected by them such as students’ GPA, course level and course type (elective vs compulsory courses). This type of research can help to have a fair investigation of the relationships between SET and these factors.

3. It is recommended to replicate this study using datasets about evaluation of courses (SEQ) using factors that are related to students, faculty members and courses.

References


**Corresponding author**

Mahmoud AlQuraan can be contacted at: mahmoud.q@yu.edu.jo